Linear classifiers are nearly optimal when hidden variables have diverse effects

Nader H. Bshouty[*]
Technion
bshouty@cs.technion.ac.il

Philip M. Long
Google
plong@google.com

Abstract

We analyze classification problems in which data is generated by a two-tiered random process. The class is generated first, then a layer of conditionally independent hidden variables, and finally the observed variables. For sources like this, the Bayes-optimal rule for predicting the class given the values of the observed variables is a two-layer neural network. We show that, if the hidden variables have non-negligible effects on many observed variables, a linear classifier approximates the error rate of the Bayes optimal classifier up to lower order terms.

1 Introduction

In many classification problems, groups of features are positively associated, even among examples of a given class. For example, when classifying news articles as to whether they are about sports or not, words about soccer tend to appear in the same articles. Similarly, biomolecules are organized into subsystems, including pathways, and diseases often coordinately affect the production rates of members of certain subsystems.

One way to model this phenomenon is to think of the class designations and feature values as being generated by a probability distribution with hidden variables [10, 3, 24, 16, 14]. In one model of this type, the class designation directly and conditionally independently affects the hidden variables, each of which in turn drives a set of observed variables (see Figure 1). Each hidden variable can be interpreted as indicating whether a group of observed variables have been collectively affected by the class of the item. For example, a hidden variable could indicate whether an article is about soccer or not. Its descendents would include words that are especially common in articles about soccer, like “corner” and “striker”. It is intuitive that the Bayes optimal classifier for a source like this is a two-layer feed-forward neural network, with the hidden layer of the neural network corresponding to the layer of hidden variables in the generative model. (We provide a proof because we are not aware of a reference for this in the literature.)

Despite this fact, for many problems clearly possessing such hierarchical structure, learning algorithms that use linear hypotheses achieve excellent, often even state-of-the-art, performance (see, e.g. [12, 19, 22, 20, 11]). This might appear paradoxical, because one might think that such algorithms must be doomed to fail because they use an inordinately limited hypothesis space.

In this paper, we show that, despite the fact that the optimal classifier has a more complex structure, a linear classifier can provide a good approximation. Here is the rough idea of the proof. Suppose that a hidden variable is binary-valued, and that its value affects many features – this is to be expected for example in text classification problems, where subtopics may have many constituent words. Then a linear combination of those observed features should be expected to be highly concentrated – the combination will be close to one value when the hidden variable takes one value, and close to another value when the hidden variable takes the other value. Consequently, this linear combination of the observed features can be viewed as an approximation to a rescaling of the hidden variable. If we replace each hidden variable with the appropriate linear combination of the observed variables that it affects, and construct a linear classifier using the replacements, the result is a linear classifier of the original features.

2 Preliminaries

2.1 The structure of the source

In a *hidden variable model*, the joint distribution of the class label Y, hidden variables $H_1, ..., H_k$, and observed variables

$$X_{01}, ..., X_{0m_0}, X_{11}, ..., X_{1m_1}, ..., X_{k1}, ..., X_{km_k}$$

(all of which take values in $\{-1, 1\}$) satisfies the conditional independence constraints shown in the Bayes Net of Figure 1. The hidden variables $H_1, ..., H_k$, along with some of the observed variables, $X_{01}, ..., X_{0m_0}$, are collectively conditionally independent given the class designation Y. Each hidden variable H_i in turn has a collection of observed variables $X_{i1}, ..., X_{im_i}$ that are conditionally independent given H_i.

We can think of the model as generating labeled random examples $(\mathbf{x}, y)$ in stages, by

- generating the class label y, and fixing it, then

[*] This work was done while Nader Bshouty was visiting Google.

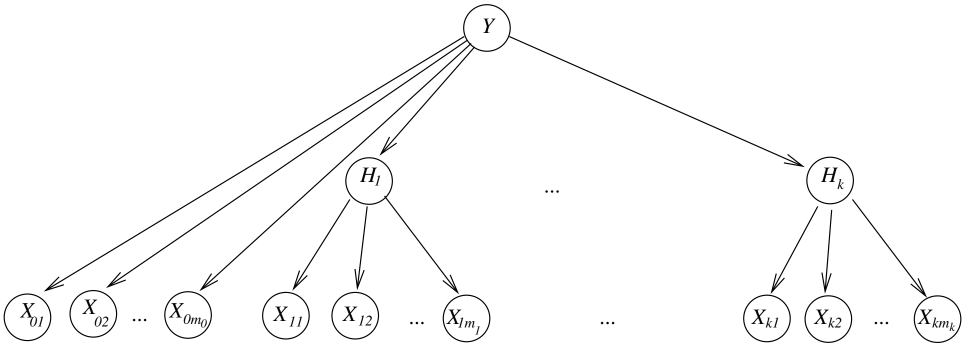

Figure 1: A probability model in which the dependence of some of the observed variables on the class designation is mediated by a layer of hidden variables.

- independently sampling the hidden variables

  $$h_1, ..., h_k$$

  and the observed variables with direct dependence on the class $x_{01}, ..., x_{0m_0}$ using the appropriate class conditional distribution, fixing them, and finally

- independently sampling the remaining observed variables

  $$x_{11}, ..., x_{1m_1}, ..., x_{k1}, ..., x_{km_k},$$

  each from the appropriate conditional distribution given the values of its parents.

Note that we may assume without loss of generality that for any indices i and j, we have

$$\mathbf{Pr}(X_{ij} = 1 | H_i = 1) > \mathbf{Pr}(X_{ij} = 1 | H_i = -1),$$

since otherwise, we could replace X_{ij} with its negation.

Definition 1 (β-effect) *We say that a hidden variable H_i β-*affects *observed variable X_{ij} if*

$$\mathbf{Pr}(X_{ij} = 1 | H_i = 1) - \mathbf{Pr}(X_{ij} = 1 | H_i = -1) > \beta.$$

To eliminate uninteresting clutter from the analysis, we will assume throughout that $\mathbf{Pr}(Y = 1) = 1/2$.

2.2 Other probability tools

Definition 2 (Total variation distance) *The* total variation distance *between probability distributions P and Q over a common domain U is $\max_{E \subseteq U} |P(E) - Q(E)|$.*

Lemma 3 (Hoeffding bound, see [17]) *Let $U_1, ..., U_\ell$ be independent zero-mean real random variables, each of which takes values in an interval of length κ. Then*

$$\mathbf{Pr}\left[\sum_{i=1}^{\ell} U_i \geq \gamma\right] \leq e^{-\frac{2\gamma^2}{\kappa^2 \ell}}.$$

3 Linear approximation

Theorem 4 *Suppose that for $\beta > 0$, each hidden variable β-affects at least m observed variables, for*

$$m = \omega\left(\frac{k \log^2(k/\mathrm{opt}) \log(1/\mathrm{opt})}{\beta^2}\right),$$

where opt *is the error rate of the Bayes optimal classifier. Suppose*

$$\mathbf{X} = (X_{01}, ..., X_{0m_0}, X_{11}, ..., X_{1m_1}, ..., X_{k1}, ..., X_{km_k})$$

are the observed variables. Then there is a linear classifier f such that

$$\mathbf{Pr}(f(\mathbf{X}) \neq Y) \leq (1 + o(1))\mathrm{opt}.$$

Proof: We will establish the stronger guarantee that the linear classifier approximates the behavior of an idealized classifier that has access to the hidden variables. The optimal classifier that uses the values $h_1, ..., h_k$ of the hidden variables $H_1, ..., H_k$ along with $x_{01}, ..., x_{km_k}$ is at least as accurate as the optimal classifier that only uses $x_{01}, ..., x_{km_k}$, since when optimizing over classifiers that have access to $h_1, ..., h_k$, one possibility is use a classifier that ignores them. The Bayes optimal classifier f_{opt}, for using a realization $\mathbf{x}$ of the tuple $\mathbf{X}$ of observed variables and a realization $\mathbf{h}$ of the tuple $\mathbf{H}$ of hidden variables, chooses $\hat{y}$ to be

$$\hat{y} = \mathrm{argmax}_y \mathbf{Pr}(Y = y | \mathbf{H} = \mathbf{h} \text{ and } \mathbf{X} = \mathbf{x}).$$

If, for each i,

$$\mathbf{X}_i = (X_{i,1}, ..., X_{i,m_i}),$$

then, since

$$H_1, ..., H_k, X_{01}, ..., X_{0m_0}$$

form a Markov blanket for Y, for any realizations

$$\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k$$

of the various groups of observed variables, we have

$$\begin{aligned} &\mathbf{Pr}(Y = y | \mathbf{H} = \mathbf{h}, \mathbf{X}_0 = \mathbf{x}_0, \mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_k = \mathbf{x}_k) \\ &= \mathbf{Pr}(Y = y | \mathbf{H} = \mathbf{h}, \mathbf{X}_0 = \mathbf{x}_0). \end{aligned}$$

Maximizing the latter is equivalent to maximizing

$$\frac{\mathbf{Pr}(Y=y|\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)}{\mathbf{Pr}(Y=-y|\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)} = \frac{\mathbf{Pr}(Y=y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)}{\mathbf{Pr}(Y=-y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)}, \qquad (1)$$

which decomposes nicely, facilitating analysis, as we will see.

The usual analysis of Naive Bayes [8] can be used to express the Bayes optimal decision rule as a linear threshold function of the variables that depend directly on Y. The odds ratio (1) can be written as follows

$$\frac{\mathbf{Pr}(Y=y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)}{\mathbf{Pr}(Y=-y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)} = \left(\left(\prod_{i=1}^{k} \frac{\mathbf{Pr}(H_i=h_i|Y=y)}{\mathbf{Pr}(H_i=h_i|Y=-y)} \right) \times \left(\prod_{j=1}^{m_0} \frac{\mathbf{Pr}(X_{0j}=x_{0j}|Y=y)}{\mathbf{Pr}(X_{0j}=x_{0j}|Y=-y)} \right) \right),$$

and a case analysis verifies that for each i, we have

$$\frac{\mathbf{Pr}(H_i=h_i|Y=y)}{\mathbf{Pr}(H_i=h_i|Y=-y)} = \exp\left(\frac{y}{2} \ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=1)}{\mathbf{Pr}(H_i=-1|Y=-1)\mathbf{Pr}(H_i=1|Y=-1)} \right) + \frac{yh_i}{2} \ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=-1)}{\mathbf{Pr}(H_i=-1|Y=1)\mathbf{Pr}(H_i=1|Y=-1)} \right) \right),$$

and similarly, for each j, we have

$$\frac{\mathbf{Pr}(X_{0j}=x_{0j}|Y=y)}{\mathbf{Pr}(X_{0j}=x_{0j}|Y=-y)} = \exp\left(\frac{y}{2} \ln\left(\frac{\mathbf{Pr}(X_{0j}=1|Y=1)\mathbf{Pr}(X_{0j}=-1|Y=1)}{\mathbf{Pr}(X_{0j}=-1|Y=-1)\mathbf{Pr}(X_{0j}=1|Y=-1)} \right) + \frac{yx_{0j}}{2} \ln\left(\frac{\mathbf{Pr}(X_{0j}=1|Y=1)\mathbf{Pr}(X_{0j}=-1|Y=-1)}{\mathbf{Pr}(X_{0j}=-1|Y=1)\mathbf{Pr}(X_{0j}=1|Y=-1)} \right) \right).$$

Thus, if for each $i \in \{1, ..., k\}$, we define

$$w_i = \frac{1}{2} \ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=-1)}{\mathbf{Pr}(H_i=-1|Y=1)\mathbf{Pr}(H_i=1|Y=-1)} \right)$$

and let

$$w_0 = \tfrac{1}{2} \textstyle\sum_{i=1}^{k} \ln\left(\frac{\mathbf{Pr}(H_i=1|Y=1)\mathbf{Pr}(H_i=-1|Y=1)}{\mathbf{Pr}(H_i=-1|Y=-1)\mathbf{Pr}(H_i=1|Y=-1)} \right)$$

and similarly, for $j \in \{1, ..., m_0\}$, let

$$v_j = \frac{1}{2} \ln\left(\frac{\mathbf{Pr}(X_{0j}=1|Y=1)\mathbf{Pr}(X_{0j}=-1|Y=-1)}{\mathbf{Pr}(X_{0j}=-1|Y=1)\mathbf{Pr}(X_{0j}=1|Y=-1)} \right)$$

and let

$$v_0 = \tfrac{1}{2} \textstyle\sum_{j=1}^{m_0} \ln\left(\frac{\mathbf{Pr}(X_{0j}=1|Y=1)\mathbf{Pr}(X_{0j}=-1|Y=1)}{\mathbf{Pr}(X_{0j}=-1|Y=-1)\mathbf{Pr}(X_{0j}=1|Y=-1)} \right),$$

then

$$\frac{\mathbf{Pr}(Y=y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)}{\mathbf{Pr}(Y=-y,\mathbf{H}=\mathbf{h},\mathbf{X}_0=\mathbf{x}_0)} = \exp\left(y\left(w_0+v_0+\mathbf{w}\cdot\mathbf{h}+\mathbf{v}\cdot\mathbf{x}_0\right)\right), \qquad (2)$$

where $\mathbf{w} = (w_1, ..., w_k)$, $\mathbf{v} = (v_1, ..., v_{m_0})$. This in turn implies that

$$f_{\mathrm{opt}}(\mathbf{x},\mathbf{h}) = \mathrm{sign}(y\left(w_0+v_0+\mathbf{w}\cdot\mathbf{h}+\mathbf{v}\cdot\mathbf{x}_0\right)).$$

(Recall that $\mathbf{x}_0$ refers the components of $\mathbf{x}$ that depend directly on the label.)

Now, let us turn to constructing the estimates of the hidden variables using the observed variables. For each i, let $\mathcal{X}_i$ consist of all indices j such that

$$\mathbf{Pr}(X_{ij}=1|H_i=1) - \mathbf{Pr}(X_{ij}=1|H_i=-1) > \beta. \qquad (3)$$

Define

$$\begin{aligned} H_i^+ &= \frac{1}{|\mathcal{X}_i|} \sum_{j\in\mathcal{X}_i} \mathbf{E}(X_{ij}|H_i=1) \\ H_i^- &= \frac{1}{|\mathcal{X}_i|} \sum_{j\in\mathcal{X}_i} \mathbf{E}(X_{ij}|H_i=-1). \end{aligned}$$

Define $\phi_i : \mathbf{R} \rightarrow \mathbf{R}$ to be the affine transformation of the real line that maps H_i^+ to 1, and H_i^- to -1; that is,

$$\phi_i(x) = \frac{2x - (H_i^+ + H_i^-)}{H_i^+ - H_i^-}.$$

For each $i > 0$, define

$$\hat{H}_i = \phi_i\left(\frac{1}{|\mathcal{X}_i|} \sum_{j\in\mathcal{X}_i} X_{ij} \right),$$

so that

$$\mathbf{E}(\hat{H}_i|H_i=h_i) = h_i. \qquad (4)$$

Now, let us analyze the accuracy of the classifier f obtained by replacing each hidden variable H_i with its estimate $\hat{H}_i$.

Our analysis will make use of the notion of a margin μ defined by

$$\mu(\mathbf{h},\mathbf{x}_0,y) = y\left(w_0+v_0+\mathbf{w}\cdot\mathbf{h}+\mathbf{v}\cdot\mathbf{x}_0\right).$$

Let us divide our analysis into the cases in which the Bayes optimal classification is made with a large margin, and cases in which its margin is small: for any $\gamma > 0$, we have

$$\begin{aligned} &\mathbf{Pr}\left[f(\mathbf{X}) \neq Y\right] \\ &= \mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right] \\ &\quad + \mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| > \gamma\right] \\ &\leq \mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right] \\ &\quad + \mathbf{Pr}\left[\mu(\mathbf{H},\mathbf{X}_0,Y) < -\gamma\right] \\ &\quad + \mathbf{Pr}\left[Y\left(\sum_{i=1}^{k} w_i\left(H_i - \hat{H}_i\right)\right) > \gamma\right] \qquad (5) \end{aligned}$$

Let us begin by working on the third term.

To control the variance of $\sum_{i=1}^{k} w_i\left(H_i - \hat{H}_i\right)$, we need to show that we can assume without loss of generality that each weight w_i is not very big. Let $\epsilon > 0$ be an error tolerance parameter that will be fixed later in the argument. Suppose we modified the source by adding a layer of hidden variables $\tilde{H}_1, ..., \tilde{H}_k$ in which

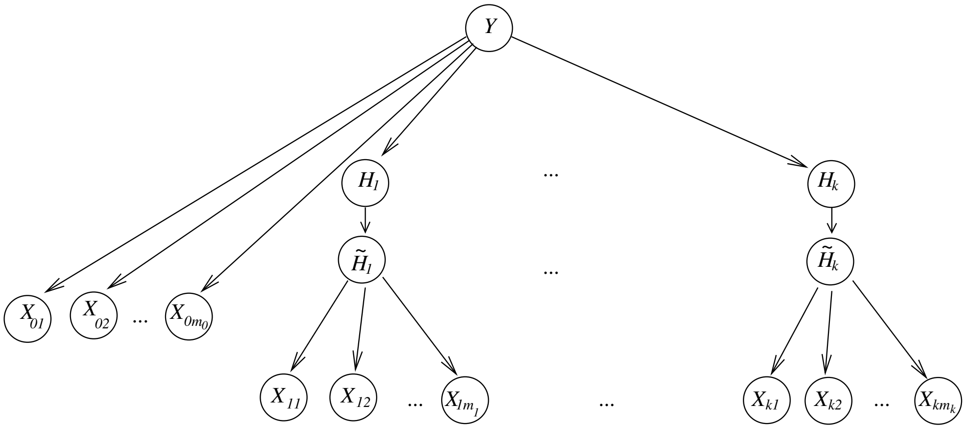

Figure 2: The dependence structure of the probability distribution used in the proof of Theorem 4.

- each $\tilde{H}_i$ was obtained by negating the value of H_i with probability ϵ/k, and
- the conditional distributions of $X_{i1}, ..., X_{im_i}$ given $\tilde{H}_i$ were the same as the old conditional distributions of $X_{i1}, ..., X_{im_i}$ given H_i.

(See Figure 2.) If we did this, the joint distribution of

$$Y, \tilde{H}_1, ..., \tilde{H}_k, X_{01}, ..., X_{0m_0}, ..., X_{k1}, ..., X_{km_k}$$

would have total variation distance at most ϵ from the distribution over

$$Y, H_1, ..., H_k, X_{01}, ..., X_{0m_0}, ..., X_{k1}, ..., X_{km_k},$$

because the probability that any of the hidden variables is flipped is at most $k(\epsilon/k) = \epsilon$. This means that the probability of error of any classifier with respect to the original source is at most ϵ more than its error probability with respect to the modified source. Furthermore,

$$Y, \tilde{H}_1, ..., \tilde{H}_k, X_{01}, ..., X_{0m_0}, ..., X_{k1}, ..., X_{km_k}$$

have the same conditional independence structure as the original source, but $\mathbf{Pr}(\tilde{H}_i = h | Y = y)$ is always in the interval $[\epsilon/k, 1 - \epsilon/k]$. The definition of w_i then implies that, by approximating the optimal classifier for the modified source, at the cost of additional error ϵ, we may assume without loss of generality that

$$W \overset{\text{def}}{=} \max_{i \geq 1} |w_i| = O(\ln(k/\epsilon)). \tag{6}$$

Furthermore, suppose that, instead of setting each of the probabilities that H_i was flipped to be ϵ/k, we instead chose each flip probability from the interval

$$[\epsilon/(2k), \epsilon/k].$$

Then the total variation distance from the original to the modified source would still be at most ϵ, and the weights would still satisfy (6). But, since each weight is a continuous function of $\mathbf{Pr}(\tilde{H}_i = h | Y = y)$, and there are only finitely many values that $\tilde{\mathbf{H}}$ and $\mathbf{X}_0$ can take, there are values in $[\epsilon/(2k), \epsilon/k]$ for the flip probabilities such that $\mathbf{Pr}(\mu(\tilde{\mathbf{H}}, \mathbf{X}_0, y) = 0) = 0$. Thus, we may also assume without loss of generality that

$$\mathbf{Pr}(\mu(\mathbf{H}, \mathbf{X}_0, y) = 0) = 0. \tag{7}$$

Now let us return to bounding the third term of (5). For some realization $y, h_1, ..., h_k$ of the label and the hidden variables, let us condition on the event that

$$Y = y, H_1 = h_1, ..., H_k = h_k. \tag{8}$$

The independence structure of the source implies that, after conditioning on (8),

$$\sum_{i=1}^{k} w_i \left(H_i - \hat{H}_i \right) = \sum_{i=1}^{k} w_i \left(h_i - \hat{H}_i \right)$$

is a sum of independent random variables. Further, by (4), $\mathbf{E}(\hat{H}_i) = h_i$, so the expectation of

$$S = \sum_{i=1}^{k} w_i \left(h_i - \hat{H}_i \right)$$

is 0. Thus, S is a sum of at least km independent random variables, each of which, by (6), and the definitions of $\mathcal{X}_i$ and $\hat{H}_i$, takes values in an interval of size at most $\frac{4W}{\beta m}$. Applying the Hoeffding bound (Lemma 3), we get

$$\begin{aligned} &\mathbf{Pr}\left[Y \left(\sum_{i=1}^{k} w_i \left(H_i - \hat{H}_i \right) \right) > \gamma \right] \\ &\leq \exp\left(\frac{-\gamma^2 \beta^2 m}{8kW^2} \right) \\ &\leq \exp\left(-\Omega\left(\frac{\gamma^2 \beta^2 m}{k \ln^2(k/\epsilon)} \right) \right). \end{aligned} \tag{9}$$

Now, let us work on the first term of (5). We can pair borderline cases with their counterparts in which the label is negated. Equation (2) implies that the two cases are nearly equally likely. Since both the linear classifier and the Bayes optimal classifier make an incorrect classification in one of the cases, the linear classifier approximates the accuracy of the Bayes optimal classifier, on average, over borderline cases. To see this logic in more detail, let us start by recalling that (7) implies that

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right]\\ &= \mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \in (0,\gamma]\right].\end{aligned}$$

Now, let us evaluate this probability with a sum over pairs of examples that differ only on the label. (Note that exactly one example from each pair will have a positive margin.)

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right]\\ &= \sum_{\mathbf{h},\mathbf{x},y:\mu(\mathbf{h},\mathbf{x}_0,y)\in(0,\gamma]} \mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=y)\\ &\qquad \times 1_{f\neq y}(\mathbf{x},y)\\ &\qquad +\mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=-y)\\ &\qquad \times 1_{f\neq -y}(\mathbf{x},-y),\end{aligned}$$

where $1_{f\neq y}$ is an indicator function for $\{(\mathbf{x},y) : f(\mathbf{x}) \neq y\}$. Since each pair of examples differs only in the label, any classifier, in particular, the linear classifier f, must classify one example of each pair correctly, thus

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right]\\ &\leq \sum_{\mathbf{h},\mathbf{x},y:\mu(\mathbf{h},\mathbf{x}_0,y)\in(0,\gamma]} \max\{\mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=y),\\ &\qquad \mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=-y)\}.\end{aligned}$$

Now, (2) immediately gives.

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right]\\ &\leq \sum_{\mathbf{h},\mathbf{x},y:\mu(\mathbf{h},\mathbf{x}_0,y)\in(0,\gamma]} e^{\gamma}\min\{\mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=y),\\ &\qquad \mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=-y)\}.\end{aligned}$$

The Bayes optimal classifier chooses $\hat{y}$ to maximize $\mathbf{Pr}(\mathbf{H} = \mathbf{h}, \mathbf{X} = \mathbf{x}, Y = \hat{y})$, so it will make an error in the opposite case. Thus

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y \text{ and } |\mu(\mathbf{H},\mathbf{X}_0,Y)| \leq \gamma\right]\\ &\leq e^{\gamma} \sum_{\mathbf{h},\mathbf{x},y:\mu(\mathbf{h},\mathbf{x}_0,y)\in[-\gamma,0)} \mathbf{Pr}(\mathbf{H}=\mathbf{h},\mathbf{X}=\mathbf{x},Y=y)\\ &= e^{\gamma}\mathbf{Pr}(\mu(\mathbf{H},\mathbf{X}_0,Y) \in [-\gamma,0)).\end{aligned}$$

Putting this together with (5) and (9), we get

$$\begin{aligned}&\mathbf{Pr}\left[f(\mathbf{X}) \neq Y\right]\\ &\leq e^{\gamma}\mathbf{Pr}\left[f_{\mathrm{opt}}(\mathbf{X},\mathbf{H}) \neq Y\right] + \epsilon\\ &\quad + \exp\left(-\Omega\left(\frac{\gamma^2\beta^2 m}{k\ln^2(k/\epsilon)}\right)\right).\end{aligned} \tag{10}$$

Now, let us show that we can choose ϵ and γ so that $\mathbf{Pr}(f(\mathbf{X}) \neq Y) = (1 + o(1))\mathrm{opt}$. Suppose $\epsilon = \mathrm{opt}^2$, so that the second term of (10) is $o(\mathrm{opt})$. Then a choice of γ that satisfies

$$\gamma = \Theta\left(\sqrt{\frac{k\log^2(k/\mathrm{opt})\log(1/\mathrm{opt})}{\beta^2 m}}\right)$$

suffices for

$$\exp\left(-\Omega\left(\frac{\gamma^2\beta^2 m}{k\ln^2(k/\epsilon)}\right)\right) = \mathrm{opt}^2$$

so that the third term of (10) is $o(\mathrm{opt})$. For such a γ,

$$m = \omega\left(\frac{k\log^2(k/\mathrm{opt})\log(1/\mathrm{opt})}{\beta^2}\right)$$

suffices for

$$e^{\gamma} = 1 + o(1),$$

completing the proof. ■

The hidden variables can afford to be much less influential if they have similar degrees of association with the class designation. This is illustrated by considering the idealized case in which all associations are equally strong.

Theorem 5 *Suppose there is* $0 < \alpha < 1/4$ *such that each hidden variable* H_i *has* $\mathbf{Pr}(H_i = y|Y = y) = 1/2 + \alpha$ *for both* $y \in \{-1, 1\}$*, and suppose* $m_0 = 0$*.*

If in addition for $\beta > 0$*, each hidden variable* β*-affects at least* m *observed variables, and* opt *is the error rate of the Bayes optimal classifier, for*

$$m = \omega\left(\frac{\log^2(1/\mathrm{opt})}{\beta^2}\right),$$

then there is a linear classifier whose error rate is

$$(1 + o(1))\mathrm{opt}.$$

Proof: First, we will modify the proof of Theorem 4 to get an upper bound, and then reason that the upper bound implies the theorem.

First, $w_i = \ln\frac{1+2\alpha}{1-2\alpha}$ for all i, so $w_i = \Theta(\alpha)$.

Thus (9) can be replaced with

$$\begin{aligned}&\mathbf{Pr}\left[Y\left(\sum_{i=1}^{k} w_i\left(H_i - \hat{H}_i\right)\right) > \gamma\right]\\ &\leq \exp\left(\frac{-c\gamma^2\beta^2 m}{\alpha^2 k}\right).\end{aligned} \tag{11}$$

for an absolute positive constant c, leading to a bound of

$$\begin{aligned}&\mathbf{Pr}(f(\mathbf{X}) \neq Y)\\ &\leq e^{\gamma}\mathrm{opt} + \exp\left(\frac{-c\gamma^2\beta^2 m}{\alpha^2 k}\right).\end{aligned} \tag{12}$$

As argued in the proof of Theorem 4, the error rate of the Bayes optimal classifier that uses only the observed variables is at least as large as the error rate of the optimal classifier that also uses the hidden variables, and, for sources considered in this theorem, the latter classifier simply takes a majority vote over the values of the hidden variables. This classifier is incorrect when a majority of the hidden variables take values different from the label. Applying the Hoeffing bound, this happens with probability $\exp(-\Omega(\alpha^2 k))$, and thus, $\alpha^2 k = O(\log(1/\mathrm{opt}))$ which implies

$$\begin{aligned}&\mathbf{Pr}(f(\mathbf{X}) \neq Y)\\ &\leq e^{\gamma}\mathrm{opt} + \exp\left(\frac{-c'\gamma^2\beta^2 m}{\log(1/\mathrm{opt})}\right).\end{aligned}$$

Here, a choice of γ that satisfies

$$\gamma = \Theta\left(\sqrt{\frac{\log^2(1/\mathrm{opt})}{\beta^2 m}}\right)$$

is sufficient for

$$\exp\left(\frac{-c'\gamma^2\beta^2 m}{\log(1/\mathrm{opt})}\right) = \mathrm{opt}^2,$$

and, for such a γ, $m = \omega\left(\frac{\log^2(1/\mathrm{opt})}{\beta^2}\right)$ suffices for $e^\gamma = 1 + o(1)$. ■

4 Bayes Optimal Models are Two-layer Neural Networks

In this section, we show that, even with further restrictions on the structure of the source, a two-layer neural network is needed to compute the exact Bayes optimal classifier.

Theorem 6 *Suppose that there are real constants $\alpha, \beta > 0$ and a positive integer m such that*

- *each H_i is independently equal to Y with probability $1/2 + \alpha$,*
- *$m_0 = 0$, and $m_i = m$ for all $i > 0$, and*
- *each X_{ij} is independently equal to H_i with probability $1/2 + \beta$,*
- *$A = \frac{1+2\alpha}{1-2\alpha}$, $B = \frac{1+2\beta}{1-2\beta}$, and, for each $i \in \{1, ..., k\}$, $s_i(\mathbf{x}) = \sum_{j=1}^m x_{ij}$.*

Then the Bayes optimal classifier is

$$h(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^k \log\left(\frac{B^{s_i(\mathbf{x})}A + 1}{B^{s_i(\mathbf{x})} + A}\right)\right). \tag{13}$$

Proof: Notice that for any $y \in \{-1, 1\}$,

$$\begin{aligned}
&\mathbf{Pr}[Y = y | (\forall i, j) X_{ij} = x_{ij}] \\
&= \frac{\mathbf{Pr}[Y = y]\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = y]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij}]} \\
&= \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = y]}{2\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij}]}
\end{aligned}$$

and therefore

$$\begin{aligned}
&\mathbf{Pr}[Y = 1 | (\forall i, j) X_{ij} = x_{ij}] \\
&> \mathbf{Pr}[Y = -1 | (\forall i, j) X_{ij} = x_{ij}]
\end{aligned}$$

if and only if

$$\begin{aligned}
&\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] \\
&> \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1].
\end{aligned}$$

Therefore, the Bayes optimal classifier gives

$$\begin{aligned}
h(\mathbf{x}) = \mathrm{sign}\,(&\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] - \\
&\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]) .
\end{aligned}$$

Since $\log$ is a monotone function we also have

$$\begin{aligned}
&h(\mathbf{x}) \\
&= \mathrm{sign}\,(\log \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] \\
&\qquad - \log \mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]) \\
&= \mathrm{sign}\left(\log \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]}\right).
\end{aligned} \tag{14}$$

Let $S_i = H_i Y$ (so that S_i that is 1 with probability $\frac{1}{2} + \alpha$ and -1 with probability $\frac{1}{2} - \alpha$), and $T_{ij} = X_{ij} H_i$ (so T_{ij} is 1 with probability $\frac{1}{2} + \beta$ and -1 with probability $\frac{1}{2} - \beta$). Now since T_{ij} and S_i are independent of Y, and, the events $[(\forall j) T_{ij} S_i = x_{ij}]$ are independent

$$\begin{aligned}
&\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1] \\
&= \mathbf{Pr}[(\forall i, j) T_{ij} S_i Y = x_{ij} | Y = 1] \\
&= \mathbf{Pr}[(\forall i, j) T_{ij} S_i = x_{ij} | Y = 1] \\
&= \mathbf{Pr}[(\forall i, j) T_{ij} S_i = x_{ij}] \\
&= \prod_{i=1}^k \mathbf{Pr}[(\forall j) T_{ij} S_i = x_{ij}].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1] \\
&= \prod_{i=1}^k \mathbf{Pr}[(\forall j) T_{ij} S_i = -x_{ij}].
\end{aligned}$$

By (14) we get

$$\begin{aligned}
&h(\mathbf{x}) \\
&= \mathrm{sign}\left(\log \frac{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = 1]}{\mathbf{Pr}[(\forall i, j) X_{ij} = x_{ij} | Y = -1]}\right) \\
&= \mathrm{sign}\left(\sum_{i=1}^k \log\left(\frac{\mathbf{Pr}[(\forall j) T_{ij} S_i = x_{ij}]}{\mathbf{Pr}[(\forall j) T_{ij} S_i = -x_{ij}]}\right)\right).
\end{aligned} \tag{15}$$

Now, since for every i,

$$\begin{aligned}
&\frac{Pr[(\forall j) T_{ij} = x_{ij}]}{Pr[(\forall j) T_{ij} = -x_{ij}]} \\
&= \prod_{j=1}^m \frac{Pr[T_{ij} = x_{ij}]}{Pr[T_{ij} = -x_{ij}]} \\
&= \prod_{j=1}^m B^{x_{ij}},
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbf{Pr}[(\forall j) T_{ij} S_i = x_{ij}] \\
&= \mathbf{Pr}[(\forall j) T_{ij} = x_{ij}]\mathbf{Pr}[S_i = 1] \\
&\qquad + \mathbf{Pr}[(\forall j) T_{ij} = -x_{ij}]\mathbf{Pr}[S_i = -1] \\
&= \left(\frac{1}{2} - \alpha\right)(\mathbf{Pr}[(\forall j) T_{ij} = x_{ij}]A \\
&\qquad + \mathbf{Pr}[(\forall j) T_{ij} = -x_{ij}]) \\
&= \left(\frac{1}{2} - \alpha\right)\mathbf{Pr}[(\forall j) T_{ij} = -x_{ij}]\left(A\prod_{j=1}^m B^{x_{ij}} + 1\right).
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{Pr}[(\forall j)T_{ij}S_i = -x_{ij}] \\
&= \mathbf{Pr}[(\forall j)T_{ij} = -x_{ij}]\mathbf{Pr}[S_i = 1] \\
&\quad +\mathbf{Pr}[(\forall j)T_{ij} = x_{ij}]\mathbf{Pr}[S_i = -1] \\
&= \left(\frac{1}{2} - \alpha\right)(\mathbf{Pr}[(\forall j)T_{ij} = -x_{ij}]A \\
&\quad +\mathbf{Pr}[(\forall j)T_{ij} = x_{ij}]) \\
&= \left(\frac{1}{2} - \alpha\right)\mathbf{Pr}[(\forall j)T_{ij} = -x_{ij}]\left(A + \prod_{j=1}^{m} B^{x_{ij}}\right).
\end{aligned}$$

Now by (15),

$$\begin{aligned}
&h(\mathbf{x}) \\
&= \operatorname{sign}\left(\sum_{i=1}^{k} \log\left(\frac{A\prod_{j=1}^{m} B^{x_{ij}} + 1}{A + \prod_{j=1}^{m} B^{x_{ij}}}\right)\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{k} \log\left(\frac{A\exp\left(\sum_{j=1}^{m}(\ln B)x_{ij}\right) + 1}{A + \exp\left(\sum_{j=1}^{m}(\ln B)x_{ij}\right)}\right)\right),
\end{aligned}$$

completing the proof. ■

One useful representation uses the following Taylor series

$$\ln x = 2\left[\left(\frac{x-1}{x+1}\right) + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \cdots\right]$$

and gives

$$h(\mathbf{x}) = \operatorname{sign}\Bigg(\sum_{i=1}^{k}\sum_{\ell=1}^{\infty}\frac{\left(2\alpha\right)^{2\ell-1}}{2\ell-1} \times \tanh^{2\ell-1}\left(\frac{1}{2}\sum_{j=1}^{m}(\ln B)x_{ij}\right)\Bigg), \qquad (16)$$

where

$$\tanh y = \frac{e^{2y} - 1}{e^{2y} + 1}.$$

The hyperbolic tangent is a standard squashing function for the hidden nodes in a two-layer neural network [9], and raising it to a positive odd power maintains the sigmoid shape. Thus the Bayes optimal classifier described in Theorem 6 can be thought of as a two-layer neural network.

The classifier of (13) approximately,

- for each i, computes an estimate V_i of H_i by taking a majority vote over $X_{i1}, ..., X_{im}$, and
- outputs a vote over V_i.

Intuitively, this is not a linear classifier, since, for example, X_{im} matters less if the value of V_i is already more-or-less determined by the values of $X_{i1}, ..., X_{i(m-1)}$. This is formalized in the following.

Theorem 7 *If $k = m = 3$, for any $\alpha > 0$, there is a value of $\beta \in (0, 1/2)$, so that the classifier h defined in (13) is not linear.*

Proof: Assume for contradiction that $\mathbf{w} \in \mathbf{R}^{km}$ is the weight vector of a linear classifier f equal to h, i.e.

$$\operatorname{sign}\left(\sum_i\sum_j w_{ij}x_{ij}\right) = h(\mathbf{x})$$

for all $\mathbf{x} \in \{-1, 1\}^{km}$.

We claim that this implies that h computes a majority vote. By symmetry, for any $\mathbf{x}$, any permutation ϕ of $\{1, ..., k\}$ and any permutation ψ over $\{1, ..., m\}$, we have

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_i\sum_j w_{ij}x_{\phi(i)\psi(j)}\right). \qquad (17)$$

In general, for real a and b, if $\operatorname{sign}(a) = \operatorname{sign}(b)$, then $\operatorname{sign}(a + b) = \operatorname{sign}(a) = \operatorname{sign}(b)$. Thus, (17) implies

$$h(\mathbf{x}) \quad = \quad \operatorname{sign}\left(\sum_\phi\sum_\psi\sum_i\sum_j w_{ij}x_{\phi(i)\psi(j)}\right).$$

This in turn implies

$$h(\mathbf{x}) \quad = \quad \operatorname{sign}\left((k-1)!(m-1)!\left(\sum_{i,j} w_{ij}\right)\sum_{i,j} x_{ij}\right)$$

because the permutations ϕ and ψ pair each weight with each feature an equal number of times. Rescaling, we get

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_{i,j} x_{ij}\right),$$

the majority function.

To arrive at a contradiction, suppose $k = m = 3$, and

$$\mathbf{x} = ((1, 1, 1), (1, -1, -1), (1, -1, -1)).$$

Note that the majority function evaluates to 1 on $\mathbf{x}$. On the other hand, using the definition in (13), we have

$$h(\mathbf{x}) = \operatorname{sign}\left(\log\left(\frac{B^3A + 1}{B^3 + A}\right) + 2\log\left(\frac{B^{-1}A + 1}{B^{-1} + A}\right)\right).$$

As β gets closer to $1/2$, B gets arbitrarily large. But

$$\begin{aligned}
&\lim_{B\to\infty}\log\left(\frac{B^3A + 1}{B^3 + A}\right) + 2\log\left(\frac{B^{-1}A + 1}{B^{-1} + A}\right) \\
&= \log A - 2\log A < 0
\end{aligned}$$

and therefore there is a value of β such that $h(\mathbf{x}) = -1$, a contradiction. ■

5 Some related work

A number of papers have considered why the Naive Bayes algorithm, which outputs a linear hypothesis, works well despite class-conditional dependencies among the features [7, 2, 23, 13]. While Naive Bayes works suprisingly well, other linear classifiers typically perform better [5, 4]. Note that Naive Bayes may not work for the sources considered in this paper.

The hidden variable model studied here is a generalization of the Neyman Model of Evolution [15]. A PAC algorithm for learning the probability distribution over the leaves for such models is known [6]. Using known tools, this algorithm can be used as a subroutine in a polynomial-time algorithm for approximating the Bayes-Optimal classifier for sources in which the class-conditional distributions are of this form [1]. The linear approximation pointed out in this paper could be a step toward a more efficient algorithm for this problem. Most examples appear to be classified correctly by a large margin, which promises to help even more.

Another related line of work is on threshold circuit complexity (see [18, 21] for surveys).

6 Conclusion

The analysis of this paper illustrates the power of linear models even in the presence of latent structure among the features. The exact mathematical statements of this paper are among many possible choices that trade off between interpretibility and coverage in different ways. For example, it would not be hard to extend the approximation to apply to sources in which fan-in is allowed. If some observed variables depend on multiple hidden variables, as long as each hidden variable has enough variables that depend on it alone, we can construct the linear approximation as a function only of the observed variables that depend on specific hidden variables.

Our analysis only demonstrates the existence of a good linear classifier, leaving the following problem open: what training algorithm is best suited to learn the parameters of a linear classifier approximating the Bayes-optimal accuracy in hidden variable models like that studied here?

Acknowledgements

We thank Fernando Pereira, Olcan Sercinoglu, Rocco Servedio and Dekang Lin for their guidance and advice.

We also thank anonymous reviewers for their careful reading of the paper, and thoughtful comments.

References

[1] S. Anoulova, P. Fischer, S. Pölt, and H. U. Simon. Probably almost Bayes decisions. *Inform. Comput.*, 129(1):63–71, August 1996.

[2] P. Bickel and E. Levina. Some theory of Fisher's linear discriminant function, 'Naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[4] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. *ICML*, pages 96–103, 2008.

[5] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. *ICML*, pages 161–168, 2006.

[6] M. Cryan, L. A. Goldberg, and P. W. Goldberg. Evolutionary trees can be learned in polynomial time in the two-state general Markov model. *SIAM J. Comput.*, 31(2):375–397, 2001.

[7] Pedro Domingos and Michael Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley, 2000.

[9] J. A. Hertz, A. Krogh, and R. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley, 1991.

[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

[11] C. J. Hsieh, K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. *ICML*, 2008.

[12] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, 1998.

[13] L. I. Kuncheva. On the optimality of naive bayes with dependent binary features. *Pattern Recognition Letters*, 27(7):830–837, 2006.

[14] H. Langseth and T. D. Nielsen. Classification using hierarchical naive bayes models. *Machine Learning*, 63(2):135–159, 2006.

[15] J. Neyman. *Molecular studies of evolution: a source of novel statistical problems*, pages 1–27. Academic Press, 1971.

[16] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *J. Comp. Sys. Sci.*, 61(1):217–235, 2000.

[17] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984.

[18] A. A. Razborov. On small depth threshold circuits. In *SWAT*, pages 42–52, 1992.

[19] R.E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, May/June 2000.

[20] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, pages 807–814, 2007.

[21] J. Sima and P. Orponen. General purpose computation with neural networks: a survey of complexity theoretic results. *Neural Computation*, 15:2727–2778, 2003.

[22] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99(10):6567–72, 2002.

[23] H. Zhang. The optimality of naive bayes. *FLAIRS*, 2004.

[24] N. L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5(6):697–723, 2004.